

# Mitigating Geospatial Knowledge Hallucination in Large Language Models: Benchmarking and Dynamic Factuality Aligning

Shengyuan Wang<sup>†</sup>, Jie Feng<sup>‡</sup>, Tianhui Liu<sup>§</sup>, Dan Pei<sup>¶</sup>, Yong Li<sup>‡</sup>

<sup>†</sup>College of AI, Tsinghua University

<sup>‡</sup>Department of Electronic Engineering, BNRist, Tsinghua University

<sup>§</sup>School of Electronic and Information Engineering, Beijing Jiaotong University

<sup>¶</sup>Department of Computer Science and Technology, Tsinghua University

## Abstract

Large language models (LLMs) possess extensive world knowledge, including geospatial knowledge, which has been successfully applied to various geospatial tasks such as mobility prediction and social indicator prediction. However, LLMs often generate inaccurate geospatial knowledge, leading to geospatial hallucinations—incorrect or inconsistent representations of geospatial information—that compromise their reliability. While the phenomenon of general knowledge hallucination in LLMs has been widely studied, the systematic evaluation and mitigation of geospatial hallucinations remain largely unexplored. To address this gap, we propose a comprehensive evaluation framework for geospatial hallucinations, leveraging structured geospatial knowledge graphs for controlled assessment. Through extensive evaluation across 20 advanced LLMs, we uncover the hallucinations in their geospatial knowledge. Building on these insights, we introduce a dynamic factuality aligning method based on Kahneman-Tversky Optimization (KTO) to mitigate geospatial hallucinations in LLMs, leading to a performance improvement of over 29.6% on the proposed benchmark. Extensive experimental results demonstrate the effectiveness of our benchmark and learning algorithm in enhancing the trustworthiness of LLMs in geospatial knowledge and reasoning tasks. Codes and data are available via <https://anonymous.4open.science/r/GeospatialHallucination-823A/>.

## 1 Introduction

Recently, large language models (LLMs), known for their excellent reasoning abilities (Wei et al., 2022) and extensive world knowledge (Yu et al., 2023; Ivanova et al., 2024), have been widely applied across various domains. The extensive geospatial knowledge embedded in LLMs has also been explored (Gurnee and Tegmark, 2023; Roberts

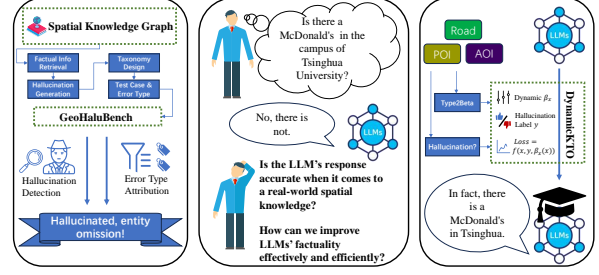


Figure 1: An overview of our work. In order to address real-world geospatial hallucination of LLMs. We propose 1) GEOHALUBENCH to detect and evaluate geospatial knowledge errors, 2) Dynamic KTO to enhance LLMs’ factuality effectively and efficiently.

et al., 2023) and successfully applied to various geospatial tasks. Gurnee et al. (Gurnee and Tegmark, 2023) and Roberts et al. (Roberts et al., 2023) demonstrate that LLMs maintain grounded geospatial knowledge that accurately reflects the real world. Leveraging this geospatial knowledge, LLM-based methods have shown promising performance in several geospatial tasks, including GeoLLM for social indicator prediction (Manvi et al., 2023), AgentMove for global mobility prediction (Feng et al., 2024b), and UrbanCLIP for robust and effective urban representation (Yan et al., 2024).

While the world knowledge embedded in LLMs has contributed to their widespread success in various applications over the past two years, researchers have identified significant errors and self-contradictions—referred to as hallucinations—in the generated results of LLMs, particularly in domain-specific areas (Huang et al., 2023; Ji et al., 2023). These hallucinations significantly affect the trustworthiness of LLMs and their performance in real-world applications. Geospatial knowledge within LLMs also exhibits notable hallucinations in practical use cases (Manvi et al., 2024; Feng et al., 2024c). Detecting and mitigating these hallucinations has become a critical problem for the develop-

ment and reliable deployment of LLMs in the real world. While various solutions have been proposed for general and domain-specific knowledge hallucinations (Zhang et al., 2024; Chen et al., 2024), systematic evaluation and mitigation of geospatial knowledge hallucinations remain largely unexplored. This is particularly challenging due to two main factors. First, geospatial data is complex and diverse, resulting in varied manifestations of associated spatial hallucinations. Second, general hallucination reduction methods often fail to account for the unique characteristics of geospatial knowledge, e.g., various elements and relations between them.

In this paper, we propose a systematic benchmark and an effective factuality aligning method to evaluate and mitigate geospatial hallucinations in LLMs. We first build *SpatialKG* to reorganize the diverse and unstructured geospatial data, and introduce a set of geospatial evaluation questions, along with a systematic taxonomy of hallucinations. Furthermore, the elements in the evaluation questions are flexible and customizable, allowing for a thorough localization of the shortcomings in geospatial hallucinations across various LLMs. After evaluating 20 advanced LLMs using our proposed benchmark, we find that most of them, especially the open-source LLMs, exhibit significant hallucinations. Based on our observations of these geospatial hallucinations, we develop DynamicKTO to effectively mitigate the geospatial hallucinations in smaller-scale, open-source LLMs by enhancing the KTO algorithm with dynamic factuality aligning. With the help of DynamicKTO, LLama3.1-8B achieves a significant performance improvement on the proposed benchmark and demonstrates competitive performance compared to the second-best model among the 20 advanced LLMs we evaluated. In summary, our contributions are threefold.

- To the best of our knowledge, we are the first to systematically evaluate and mitigate the geospatial hallucinations in LLMs.
- We have developed a comprehensive benchmark to assess geospatial hallucinations and analyze the performance of 20 advanced LLMs within this framework.
- We propose DynamicKTO, which extends the KTO by incorporating dynamic factuality aligning to account for the diversity and heterogeneity of geospatial knowledge and data.
- Extensive experiments on GEOHALUBENCH

and DynamicKTO demonstrate the effectiveness of our proposed framework in evaluating and mitigating geospatial hallucinations.

## 2 Methods

### 2.1 *SpatialKG*: Structured Geospatial Knowledge Organization

As a reliable and informative knowledge base is essential for the definition, detection, and mitigation of the hallucination problem. We construct a high-quality knowledge graph called *SpatialKG* based on previous work (Liu et al., 2023b). We design a new schema (the high-level structure of KG, including the types of entities and relations) in order to capture fundamental elements in the urban environment and to cover most important relations for geospatial cognition.

In *SpatialKG*, fundamental **entities** include Point of Interest (POI), Area of Interest (AOI) and Road as basic elements describing urban and rural structures. Based on the types of entity in *SpatialKG*, we conclude the typical and important **relations** to describe the spatial connections between entities as follows: POI-LocateAt-AOI, POI-Near-POI, AOI-Near-AOI, AOI-ConnectTo-Road, and Road-Intersect-Road. Mastering the real-world knowledge does not only imply the memory of existing entities’ names, but also refers to the capability of recognizing their important attributes. We select the following **attributes** and link them to the entities in *SpatialKG*. For POI, address and category are basic information considered. For regions, attributes include area and type of land use (industrial, residential, etc.). As for Road, we select the length of a road as its attribute discussed.

*SpatialKG* is automatically constructed from OpenStreetMap<sup>1</sup> and Foursquare’s Open Source Places<sup>2</sup>, which are updating and high-quality city data sources. We design a pipeline to examine and filter the original data for quality control.

### 2.2 GEOHALUBENCH: Geospatial Hallucination Benchmarking

With adequate knowledge from *SpatialKG*, we can classify and benchmark LLM’s geospatial hallucination of real world.

<sup>1</sup><https://www.openstreetmap.org>

<sup>2</sup><https://opensource.foursquare.com/os-places/>

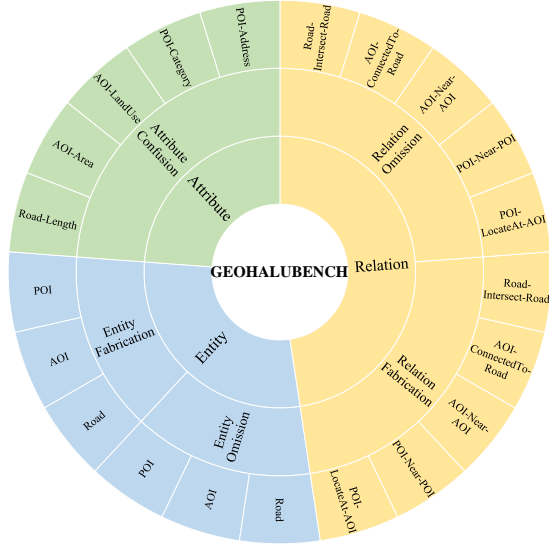


Figure 2: An illustration of the hierarchy and composition of GEOHALUBENCH.

### 2.2.1 GEOHALUBENCH Composition

In this section, we introduce the overall composition of GEOHALUBENCH. As shown in Figure 2, we organize data from *SpatialKG* into a systematic benchmark with 3 first-level categories and 5 second-level categories. GEOHALUBENCH consists of several branches, each corresponding to the regions of a city. Each city includes 21 tasks with a total of 2,100 instances.

### 2.2.2 Spatial World Hallucination Taxonomy

The concept of hallucination traces its roots to the fields of pathology and psychology (Macpherson and Platchias, 2013). Within the realm of NLP, hallucination is typically referred to as a phenomenon in which the generated content appears nonsensical or unfaithful to the provided source content (Filippova, 2020; Maynez et al., 2020). However, existing studies on hallucinations in LLMs typically define hallucination as the generation of incorrect content in terms of factuality or faithfulness (Maynez et al., 2020; Ji et al., 2023; Huang et al., 2023; Xu et al., 2024).

However, these broad definitions can be vague and insufficient for guiding further and in-depth research for specific fields. To address this gap, we propose a taxonomy for spatial world knowledge hallucination, grounded in knowledge structure. Briefly speaking, geospatial hallucination refers to fabrication of non existing geospatial entities or relations, omission of actual existences, and confusion of their attributions in this work, which is an important subset of hallucination. A more

detailed taxonomy within the concept of geospatial hallucination.

**Entity-wise.** In this first-level category, we consider the entities in the scenario of hallucination. There are two subcategories that detect different types of hallucinations. **1) Entity Fabrication:** LLMs will fabricate facts that do not exist actually. **2) Entity Omission:** LLMs will forget factually existing entities.

**Relation-wise.** Another first-level category roots from the relation among entities. The two subcategories define and examine the hallucination types in terms of mutual relations between entities. **1) Relation Fabrication:** This type of hallucination refers to the error of fabricating a factual inaccurate relation between entities. **2) Relation Omission:** It is a type of hallucination that omits actual relations in the real world.

**Attribute-wise.** This category represents a common type of the geospatial hallucination in the real world. **1) Attribute Confusion.** Even the knowledge about some entities is reliable, there may be errors about its attribute, like category, length, area, etc. in the field of geospatial cognition.

### 2.2.3 GEOHALUBENCH Construction

Every testing sample in GEOHALUBENCH contains a multiple choice question, one reference answer, and a mapping from options to hallucination types. The detailed construction pipeline is described below.

**Factual Information Retrieval.** Benchmarking hallucination is based on reliable factual data source. According to the proposed taxonomy of geospatial hallucination, we sample from *SpatialKG* in predefined patterns for entity, entity-entity relation, or entity-attribute. They are used as ground truths.

**Hallucination Generation.** In each test case, we curate distracting options corresponding to different error types to detect the hallucination type of the testee. These distracting options need hallucination data with given hallucination type. For Entity Fabrication, we first construct non-existing but plausible entities by instructing Meta-Llama-3.1-405B-Instruct. Then, the actual entities in reality are filtered out through comparing with *SpatialKG*. For Relation Fabrication category, irrelevant relations within *SpatialKG* are created and introduced as hallucinated information. Entity Omission and Relation Omission utilize a void option (None of the others) as the negative option. As for Attribute

Confusion, we randomly select values of attribute that is not close to the accurate value than a set threshold.

Figure 3: An example case of GEOHALUBENCH, where the multiple-choice options are labeled with distinct colors, and the corresponding hallucination types are highlighted in corresponding colors for clarity.

**Test Question**  
 Here is a multiple-choice question:  
 Which of the following is a point of interest in Beijing?  
 A. Silver Spoon Cafe  
 B. Haidian Library  
 C. None of the other options  
 Please select from A, B, C. Output your answer directly

---

**Hallucination Type**  
 Hallucinated, Entity Fabrication  
 Factual  
 Hallucinated, Entity Omission

After factual information retrieval and hallucination generation, the data are transformed into a multiple-choice question, where each option corresponds to a specific type of hallucination or a non-hallucination response. An example is demonstrated in Figure 3.

## 2.3 DynamicKTO: Optimization for Hallucination Mitigation

### 2.3.1 Kahneman-Tversky Optimization

Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) is a human-aware loss that directly maximizes the utility of generations inspired by a Kahneman-Tversky model of human utility. Inventors have shown that it matches or exceeds the performance of preference-based methods like Direct Preference Optimization (DPO) (Rafailov et al., 2023) with a more flexible data requirement.

### 2.3.2 DynamicKTO

In standard KTO, a hyperparameter  $\beta \in \mathbb{R}^+$  is introduced to the value function as a control of risk aversion, which serves a similar effect as  $\beta$  in the DPO loss, controlling how far  $\pi_\theta$  drifts from  $\pi_{\text{ref}}$ . However, a fixed beta means the same risk management strategy throughout the dataset, which is not appropriate with varying training data. If the answer to a task is relative easy and fixed, a higher  $\beta$  will encourage a closer generation with training samples to avoid risk, resulting in a better perfor-

mance, vice versa when answers are less certain. A single, unified  $\beta$  value is inadequate for addressing the diverse tasks involved in hallucination mitigation, which is further illustrated by an additional theoretical analysis in Appendix A.2.

Therefore, we propose DynamicKTO, an improved version of the Kahneman-Tversky Optimization (KTO) algorithm for hallucination mitigation, where the hyperparameter  $\beta$  is dynamically adjusted. The dynamic  $\beta$  is a function of the training sample’s feature, allowing more flexible adaptation during the optimization process. The loss function is as follows:

$$L_{\text{DynamicKTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x, y \sim D} [\lambda_y - v(x, y)],$$

where

$$r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)},$$

$$z_0 = \text{KL}(\pi_\theta(y'|x) \parallel \pi_{\text{ref}}(y'|x)),$$

$$\beta(x) = \text{Type2Beta}(x),$$

$$v(x, y) = \begin{cases} \lambda_D \sigma(\beta(x)(r_\theta(x, y) - z_0)) & \text{if } y \sim y_{\text{desirable}}|x, \\ \lambda_U \sigma(\beta(x)(z_0 - r_\theta(x, y))) & \text{if } y \sim y_{\text{undesirable}}|x. \end{cases}$$

For geospatial hallucination mitigation,  $\beta$  is adjusted to 0.1, 0.3, and 0.5 for Entity, Relation, and Attribute respectively.

## 3 Experiments

### 3.1 Benchmarking Spatial Hallucination: GEOHALUBENCH

We systematically evaluate representative LLMs on their geospatial hallucination situation with GEOHALUBENCH. All evaluations are conducted under zero-shot setting with each model’s default prompts. We use the greedy decoding strategy for all LLMs to ensure reproducibility. Following standard practices, Accuracy based on pattern matching is used as the primary metric.

**Results: Hallucination Level.** Main results of Beijing are shown in Table 1. The general performance results are relatively low, revealing significant geospatial hallucinations among different LLMs, even state-of-the-art ones. Parameter size plays a crucial role in determining the level of hallucination. For instance, within the Qwen family, larger models generally exhibit higher performance when handling spatial world knowledge. However, it is noteworthy that two Qwen2.5 models, with parameter sizes under 3B, perform exceptionally well in this task despite their smaller scale. This suggests that spatial factual tests present a unique challenge compared to other tasks, offering potential for improvement even with more limited model



Table 1: Results of GEOHALUBENCH on Beijing. For open-source LLMs, the results are presented in descending order of model size. The reported accuracy represents the macro average across the three dimensions in the "Overall" category and the micro average for all other categories. "Ranking" indicates the model’s position among the tested LLMs.

Model	License	Size	Entity	Relation	Attribute	Overall	Ranking
Gemini-2.0-flash	Proprietary	-	<b>0.4767</b>	<b>0.4936</b>	<b>0.5400</b>	<b>0.5034</b>	1
Qwen-max-2025-01-25	Proprietary	-	0.4600	0.4864	0.5160	0.4875	2
Qwen-plus-2025-01-25	Proprietary	-	0.4183	0.4840	0.5360	0.4794	3
GPT-4o-mini	Proprietary	-	0.4467	0.4520	0.4640	0.4542	6
GPT-4o	Proprietary	-	0.4083	0.4344	0.4680	0.4369	8
DeepSeek-V3	Open	671B	0.4667	0.4528	0.4920	0.4705	4
Llama-3.1-405B-Instruct	Open	405B	0.3683	0.4216	0.4480	0.4126	14
Qwen2.5-72B-Instruct	Open	72B	0.3983	0.4304	0.5360	0.4549	5
Llama-3.3-70B-Instruct	Open	70B	0.3933	0.3952	0.4800	0.4228	11
Llama-3.1-70B-Instruct	Open	70B	0.3717	0.4016	0.4520	0.4084	15
Qwen2.5-32B-Instruct	Open	32B	0.4100	0.3704	0.5040	0.4281	10
Mistral-Small-24B-Instruct-2501	Open	24B	0.2867	0.1920	0.3280	0.2689	20
Phi-4	Open	14B	0.4317	0.4072	0.5040	0.4476	7
Qwen2.5-14B-Instruct	Open	14B	0.3950	0.3920	0.4800	0.4223	12
Llama-3.1-8B-Instruct	Open	8B	0.4183	0.3752	0.2400	0.3445	19
Qwen2.5-7B-Instruct	Open	7B	0.3550	0.3056	0.5040	0.3882	16
Qwen2.5-3B-instruct	Open	3B	0.3233	0.3272	0.4600	0.3702	17
Qwen2.5-1.5B-instruct	Open	1.5B	0.3517	0.4328	0.4600	0.4148	13
Qwen2.5-0.5B-instruct	Open	0.5B	0.4400	0.4552	0.3920	0.4291	9
Random	-	-	0.4100	0.3880	0.2680	0.3553	18

sizes. The top three LLMs in GEOHALUBENCH are all proprietary models, which require substantial resources for both training and inference. Nevertheless, several open-sourced models have proven competitive, even outperforming GPT-4o.

**Results: Hallucination Types.** The distribution of hallucination types of Beijing is presented in Figure 4. The behavior of hallucination varies across different LLMs, depending on factors such as model size and training pipelines. With a few exceptions, most of the LLMs tested are able to follow the instructions in GEOHALUBENCH relatively accurately. In general, the ratio of Omission to Fabrication is notably higher than expected, suggesting that the primary cause of geospatial hallucination is a lack of knowledge, rather than the generation of non-factual content. In contrast to other model families, the LLaMA series demonstrates a pronounced tendency to fabricate assertions, a tendency that diminishes as the model’s parameter size increases. The Qwen family exhibits a spindle-shaped pattern with respect to Entity/Relation Omission as model size scales, with models ranging from 3B to 32B being more prone to omitting real-world information. The incidence of hallucinations involving attribute

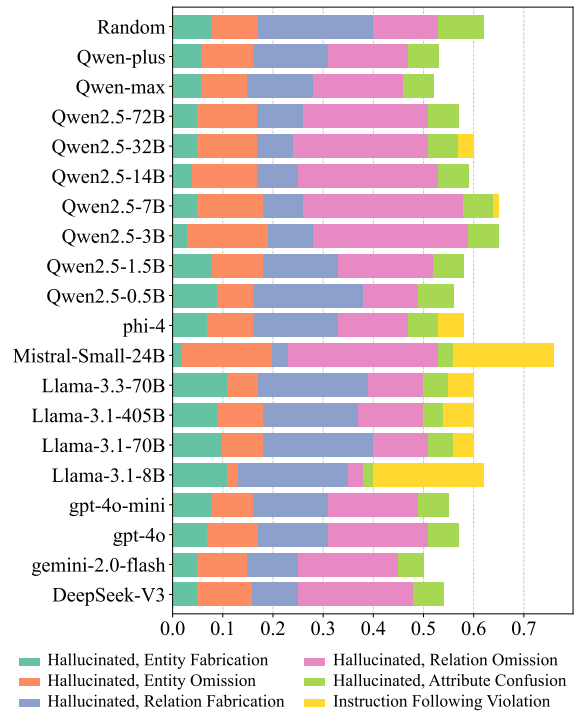


Figure 4: Distributions of Hallucination Types and Instruction Following Violation of different LLMs. Models are listed at name order. All the LLMs are chat models instruction-tuned by their developers.

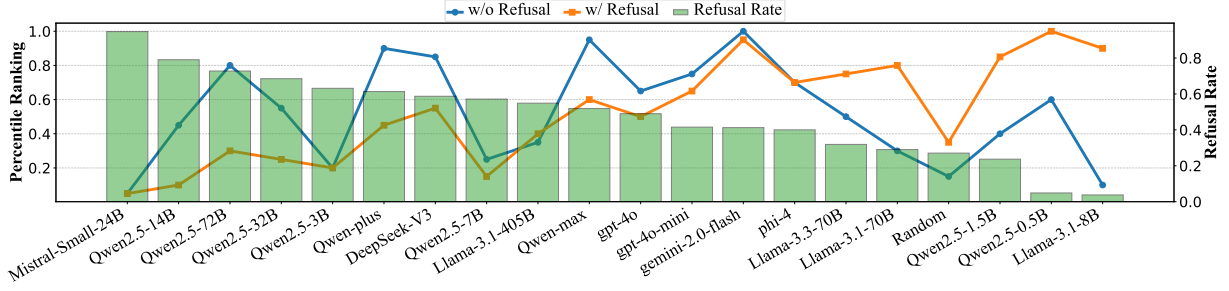


Figure 5: The introduction of the Refuse/Abstain option alters the ranking of LLMs. The line chart illustrates the percentile rankings of LLMs on GEOHALUBENCH (without refusal) and GEOHALUBENCH-Abstain (with refusal). Higher points indicate superior rankings and better performance on the respective benchmark.

Table 2: Representative results on Beijing, London, and New York. All the LLMs are chat models instruction-tuned by their developers. Results are sorted by model name.

Model	Beijing				London				New York			
	Entity	Relation	Attribute	Overall	Entity	Relation	Attribute	Overall	Entity	Relation	Attribute	Overall
DeepSeek-V3	0.4667	0.4528	0.4920	<b>0.4705</b>	0.4300	0.3904	0.4080	0.4095	0.5150	0.4336	0.4760	0.4749
GPT-4o-mini	0.4467	0.4520	0.4640	0.4542	0.4767	0.4808	0.4360	<b>0.4645</b>	0.5800	0.4784	0.4680	0.5088
Llama-3.3-70B	0.3933	0.3952	0.4800	0.4228	0.4683	0.4304	0.4720	0.4569	0.5767	0.4472	0.5200	<b>0.5146</b>
Llama-3.1-70B	0.3717	0.4016	0.4520	0.4084	0.4533	0.3976	0.3920	0.4143	0.5150	0.4416	0.4920	0.4829
Llama-3.1-8B	0.4183	0.3752	0.2400	0.3445	0.3400	0.2432	0.2640	0.2824	0.4333	0.3360	0.2680	0.3458
Qwen2.5-72B	0.3983	0.4304	0.5360	0.4549	0.4100	0.3496	0.4400	0.3999	0.5100	0.4256	0.5240	0.4865
Qwen2.5-7B	0.3550	0.3056	0.5040	0.3882	0.3317	0.2856	0.4080	0.3418	0.4033	0.3384	0.5160	0.4192
Random	0.4100	0.3880	0.2680	0.3553	0.4317	0.3704	0.3040	0.3687	0.4100	0.3928	0.2280	0.3436

confusion remains relatively high and stable across models, suggesting that this issue either receives limited attention or is inherently difficult to address within current LLM practices.

**Results: Multiple Regions.** The results of GEOHALUBENCH for Beijing, London, and New York are presented in Table 2. Due to space limitations, only representative models and results are included. Across all three cities, the overall performance is relatively low, highlighting the challenges LLMs face in handling global spatial knowledge. Although the absolute accuracy varies between cities, the performance rankings of different LLMs remain highly consistent across all three locations. This consistency demonstrates the robustness and generalizability of the GEOHALUBENCH design. Notably, hallucinations related to New York are systematically less frequent than those for Beijing or London, suggesting a general bias in the world knowledge that current LLMs tend to acquire.

Apart from representative developed cities, we also explore geospatial hallucination situation in underrepresented regions. As shown in Table 9, the lower performance results demonstrate that LLMs have more tendency to hallucination in less represented areas, showcasing LLMs’ less knowledge on these regions.

**Extension: Abstain to Answer.** For humans, it is

natural to abstain from or refuse to answer a question when faced with a knowledge gap. However, this issue becomes more complex when applied to LLMs. While abstaining from answering can help avoid factual errors, it also renders LLMs unreliable or unhelpful as knowledge sources. To evaluate LLMs’ hallucination behavior more comprehensively, we incorporate the act of abstaining or refusing to answer into our analysis. We expand GEOHALUBENCH to GEOHALUBENCH-Abstain by adding a new option of "Cannot Determine" and conduct an evaluation, with results presented in Figure 5. Refusing to answer world knowledge questions is a common behavior among LLMs, though the refusal rates vary across different models. This behavior significantly influences performance rankings on GEOHALUBENCH-Abstain, showing a noticeable negative correlation between refusal rate and performance. For instance, LLMs with lower rate of refusal like Llama-3.1-70B, Llama-3.1-8B, Qwen2.5-0.5B or Qwen2.5-1.5B have a huge improvement in ranking. On the other hand, decrease of ranking is generally associated with a high denial rate. LLMs have different strategies facing the option of abstain originated from different pre-training or post-training process. When an LLM overuses abstention, it misses opportunities to provide correct answers. Thus, strictly prohibiting or

enforcing abstention is neither practical nor reasonable. Effective training should balance precision and recall when teaching LLMs to abstain.

### 3.2 Mitigating Hallucination with DynamicKTO

Our evaluation on GEOHALUBENCH shows the weakness of current LLMs about the topic of geospatial knowledge in the real world. This situation calls for an effective method to inject LLMs with knowledge about the real world and discourage them from generating hallucinated contents. We implement DynamicKTO and validate its strength compared with existing training methods. Furthermore, we prove that DynamicKTO is not destructive to LLM’s original capabilities. Finally, we utilize DynamicKTO with supervised fine-tuning (SFT) to build a more factual model with more accurate world knowledge and capable of various urban spatial tasks.

**Datasets and Baselines.** Due to space limit, please refer Appendix A.11 for details about datasets and baselines used in experiments.

Table 3: DynamicKTO outperforms other fine-tuning methods in mitigating hallucinations. The metric used is accuracy, where the accuracy value represents the macro average across three dimensions for the "Overall" category, and a micro average for the remaining categories. Models are evaluated on GEOHALUBENCH-Abstain. The last two rows illustrate the relative improvements.

Method	Entity	Relation	Attribute	Overall
Llama3.1-8B	0.4050	0.3624	0.2840	0.3505
+SFT	0.3833	0.3568	0.2960	0.3454
+DPO	0.4183	0.3768	0.2920	0.3624
$\beta = 0.1$	0.4333	0.3912	0.3000	0.3748
+KTO $\beta = 0.3$	0.4300	0.3832	0.2880	0.3671
$\beta = 0.5$	0.4167	0.3736	0.2920	0.3608
+SimPO	0.4367	0.3928	0.3120	0.3805
+ORPO	0.4383	<b>0.4320</b>	0.3680	0.4128
+DynamicKTO	<b>0.5717</b>	0.4256	<b>0.4600</b>	<b>0.4858</b>
vs. not fine-tuned	+41.16%	+17.44%	+61.97%	+38.61%
vs. best KTO	+31.94%	+8.79%	+53.33%	+29.60%

**Results on GEOHALUBENCH.** Table 3 illustrates the advantage of DynamicKTO in mitigating hallucinations. The model trained with DynamicKTO achieves state-of-the-art (SOTA) performance, or is very close to it, across all three dimensions. When compared to the base model without fine-tuning (Llama3.1-8B-Instruct), a significant reduction in hallucinations is observed. While other fine-tuning or alignment methods generally offer improvements in hallucination mitigation, DynamicKTO outperforms them significantly, as evidenced by the results on GEOHALUBENCH. Furthermore, we

Table 4: DynamicKTO does not cause catastrophic interference with the model’s general capabilities. We utilize three renowned general benchmarks: IFEval (Zhou et al., 2023), BBH (Suzgun et al., 2022), and MMLU (Hendrycks et al., 2020).

Benchmark	Metric	Llama3.1-8B	+KTO	+ORPO	+DynamicKTO
IFEval	Accuracy	79.55	77.98	77.89	78.30
BBH	Score	44.33	43.82	42.84	43.43
MMLU	Accuracy	69.17	68.99	68.17	63.84

have tested various values of  $\beta$  in KTO to isolate the effect of hyperparameters. The improvement of DynamicKTO over the best KTO result confirms that the performance boost is not merely due to trivial hyperparameter optimization, but rather stems from the dynamic design of DynamicKTO.

**Results on General Benchmarks.** Concurrently, DynamicKTO effectively reduces hallucinations while maintaining the model’s general capabilities, as evidenced by the results presented in Table 4. Experiments demonstrate that DynamicKTO is as safe as existing methods. For two of three benchmarks, the degradation of performance is less with DynamicKTO compared with KTO or ORPO. As for MMLU, the drop of performance is around 5%. In summary, the model maintains its instruction following ability and general knowledge after trained with DynamicKTO.

**Model Generalizability.** As an alignment optimization method, DynamicKTO is model-agnostic and effective across various models in mitigating geospatial hallucinations. Comparison of different models using DynamicKTO is provided in Table 5.

### FactualCityGPT.

Recently, there has been growing interest in enhancing LLMs with real-world cognition and intelligence. Previous work of CityGPT has proposed a LLM with enhanced capabilities on understanding urban space and solving related tasks. However, we observe severe hallucination of spatial knowledge after reproducing and evaluating it, as shown in Table 3. To reduce hallucinations, improve its reliability and further enhance LLMs’ ability to handle urban tasks, we apply DynamicKTO on CityGPT. As shown in Table 6, the new SFT+DynamicKTO model is still capable of urban spatial tasks, and hallucinate less. Compared to other alignment algorithm baselines, DynamicKTO have significant advantages as well.

## 4 Related Work

**Geospatial Knowledge in LLMs** Trained on large-scale text corpora, LLMs have acquired exten-

Table 5: Effects of DynamicKTO with different base models. Both base models are chat models instruction-tuned.

Model	Before Fine-tuning				KTO ( $\beta = 0.1$ )				DynamicKTO			
	Entity	Relation	Attribute	Overall	Entity	Relation	Attribute	Overall	Entity	Relation	Attribute	Overall
Llama3.1-8B	0.4050	0.3624	0.2840	0.3505	0.4333	0.3912	0.3000	0.3748	0.5717	0.4256	0.4600	<b>0.4858</b>
Qwen2.5-7B	0.2533	0.1408	0.1480	0.1807	0.2600	0.1400	0.1600	0.1867	0.2967	0.2040	0.2280	<b>0.2429</b>

Table 6: We build Factual-CityGPT, an urban spatial LLM with reduced hallucinations. This table presents its performance on hallucination mitigation (GEOHALUBENCH) and urban spatial tasks (CityEval). CI refers to City Image, US to Urban Semantics, SRR to Spatial Reasoning Route, and SRNR to Spatial Reasoning NoRoute.

Model	Hallucination				CityEval			
	Entity	Relation	Attribute	Overall	CI	US	SRR	SRNR
CityGPT-Llama3.1-8b	0.3833	<b>0.3568</b>	0.296	0.3454	0.5492	<b>0.7000</b>	<b>0.8440</b>	<b>0.6460</b>
Factual-CityGPT-Llama3.1-8b	<b>0.5917</b>	0.3456	<b>0.3760</b>	<b>0.4378</b>	<b>0.5569</b>	0.6933	0.8040	0.6160

sive world knowledge (Ivanova et al., 2024; Yu et al., 2023), including global geospatial information (Roberts et al., 2023; Gurnee and Tegmark, 2023). This embedded geospatial knowledge has inspired the potential application of LLMs in various knowledge-intensive geospatial tasks, such as global geospatial prediction (Manvi et al., 2023)—including health, education, and poverty level estimation—mobility prediction (Wang et al., 2023; Feng et al., 2024b) using text-based addresses, and urban task planning (Jiang et al., 2024). However, due to the limitations of online corpora in capturing real-world information, researchers have explored various fine-tuning methods to enhance LLMs’ geospatial knowledge, such as CityGPT (Feng et al., 2024a) and LAMP (Balsebre et al., 2024). Unlike these studies, which focus on leveraging LLMs’ geospatial knowledge for specific tasks, our work is the first to systematically evaluate geospatial knowledge hallucinations and propose an effective mechanism to mitigate them.

**Hallucination Evaluation of LLMs.** The hallucination problem in LLMs has been widely studied (Huang et al., 2023; Ji et al., 2023), with numerous evaluation benchmarks (Bao et al., 2024; Li et al., 2023a; Liu et al., 2023a) and training methods (Ethayarajh et al., 2024; Wu et al., 2024) proposed to address general hallucination issues in LLMs. For hallucination evaluation and detection, Niels et al. (Mündler et al., 2023) investigate the problem of self-contradiction, while Manakul et al. (Manakul et al., 2023) introduce SelfCheck-GPT, a simple sampling-based method to detect hallucinations. Min et al. (Min et al., 2023) propose FactScore to identify hallucinations in long-form text generated by LLMs. Recently, Ribeiro et al. (Ribeiro et al., 2022) and Sansford et al. (Sansford et al., 2024) introduced KG-based frameworks

for hallucination detection and evaluation.

**Hallucination Reduction of LLMs.** To mitigate hallucinations in LLMs, one simple approach involves using retrieval-augmented generation (RAG) methods with external knowledge bases during generation (Lewis et al., 2020). However, RAG-based methods are resource-intensive, requiring a large number of tokens and time during inference. As a result, researchers have continued to explore more efficient methods (Zhang et al., 2024; Tian et al., 2024; Chen et al., 2024) to effectively mitigate hallucinations in various domains. For instance, Zhang (Zhang et al., 2024) proposes KnowPAT, which constructs a preference set and introduces a new alignment objective for service and urology. Chen et al. (Chen et al., 2024) propose HALC, a robust auto-focal grounding mechanism for reducing object hallucinations in vision-language models (VLMs), while Tian et al. (Tian et al., 2024) develop FactTune, a fine-tuning method aimed at reducing hallucinations in biographies and medical queries. In contrast to these works focused on general knowledge and domain-specific hallucination evaluation and mitigation, our research specifically targets the evaluation and mitigation of geospatial knowledge hallucinations in LLMs.

## 5 Conclusion

We propose a framework to systematically evaluate and mitigate geospatial hallucinations in LLMs. Using a dedicated taxonomy and controllable evaluation design, we assess 20 advanced LLMs and provide a comprehensive analysis. To improve performance, we enhance the KTO algorithm with dynamic factuality aligning, accounting for geospatial data diversity. With DynamicKTO, smaller open-source models achieve competitive results against top-performing LLMs in hallucination mitigation.



## 6 Limitations

**To Cover Broader World Knowledge.** While GEOHALUBENCH currently includes three globally diverse cities—spanning from Asia to America—there are still many regions that should be considered to fully evaluate a world model. Beijing, London, and New York are prosperous cities, but other underdeveloped areas are often less represented, where LLMs may possess less knowledge and exhibit more hallucinations. Additionally, despite *SpatialKG* captures key elements of urban space, it can be further enriched with additional information, such as the opening times of POIs, road speed limits, and more. Multi-modal data, including remote sensing or street image services, could also serve as valuable sources of world knowledge. Integrating such data into *SpatialKG* and GEOHALUBENCH could provide a more comprehensive understanding and evaluation of the world, beyond just geospatial data.

**To Generalize DynamicKTO to More Tasks.** We have proposed DynamicKTO and demonstrated its superiority in mitigating spatial knowledge hallucination, but further testing is required to benchmark DynamicKTO’s performance on other general alignment tasks. Utilizing DynamicKTO to enhance LLM’s factuality in other domain is also promising. For instance, in the area of law, medicine, finance, etc. there is a strong need for LLMs with less hallucination.

**To Explore Various Behaviors.** Teaching an LLM to say "I don’t know" is a very exciting and intriguing research question, closely tied to the topic of exploring the knowledge boundaries of LLMs. We explore LLMs behavior of abstaining from answering geospatial knowledge questions, revealing its complexity and potential. Future work should focus on finding a balance between precision and recall when training LLMs to abstain from providing answers.

## 7 Ethics Statement

All the data used in our benchmark and training algorithms come from publicly available sources, including OpenStreetMap<sup>3</sup> and Foursquare<sup>4</sup>. We adhere to their respective licenses and have made the code and data available for public access.

<sup>3</sup><https://www.openstreetmap.org>

<sup>4</sup><https://opensource.foursquare.com/os-places/>

## References

- Pasquale Balsebre, Weiming Huang, and Gao Cong. 2024. Lamp: A language model on the map. *arXiv preprint arXiv:2403.09059*.
- Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. *HHEM-2.1-Open*.
- Cody Blakeney, Mansheej Paul, Brett W Larsen, Sean Owen, and Jonathan Frankle. 2024. Does your data spark joy? performance gains from domain upsampling at the end of training. *arXiv preprint arXiv:2406.03476*.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning*.
- LMDeploy Contributors. 2023a. Lmdeploy: A toolkit for compressing, deploying, and serving llm. <https://github.com/InternLM/lmdeploy>.
- OpenCompass Contributors. 2023b. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.
- Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. 2024a. Citygpt: Empowering urban spatial cognition of large language models. *arXiv preprint arXiv:2406.13948*.
- Jie Feng, Yuwei Du, Jie Zhao, and Yong Li. 2024b. Agentmove: Predicting human mobility anywhere using large language model based agentic framework. *arXiv preprint arXiv:2408.13986*.
- Jie Feng, Jun Zhang, Junbo Yan, Xin Zhang, Tianjian Ouyang, Tianhui Liu, Yuwei Du, Siqi Guo, and Yong Li. 2024c. Citybench: Evaluating the capabilities of large language model as world model. *arXiv preprint arXiv:2406.13945*.
- Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870.
- Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

- Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Yue Jiang, Qin Chao, Yile Chen, Xiucheng Li, Shuai Liu, and Gao Cong. 2024. Urbanllm: Autonomous urban activity planning and management with large language models. *arXiv preprint arXiv:2406.12360*.
- Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583.
- Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2023. Realtime qa: What’s the answer right now? *Advances in neural information processing systems*, 36:49025–49043.
- Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. 2023. Certifying llm safety against adversarial prompting. In *First Conference on Language Modeling*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Bruce W Lee, Hyunsoo Cho, and Kang Min Yoo. 2024. *Instruction tuning with human curriculum*. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1281–1309, Mexico City, Mexico. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Danny D. Leybzon and Corentin Kervadec. 2024. *Learning, forgetting, remembering: Insights from tracking LLM memorization during training*. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 43–57, Miami, Florida, US. Association for Computational Linguistics.
- Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.
- Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. 2023b. *GeoLM: Empowering language models for geospatially grounded language understanding*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, Singapore. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Yu Liu, Jingtao Ding, Yanjie Fu, and Yong Li. 2023b. Urbankg: An urban knowledge graph system. *ACM Transactions on Intelligent Systems and Technology*, 14(4):1–25.
- Fiona Macpherson and Dimitris Platchias. 2013. *Hallucination: Philosophy and psychology*. MIT Press.
- Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.
- Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.
- Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

- Yu Meng, Mengzhou Xia, and Danqi Chen. 2025. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66.
- Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.
- Leonardo FR Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. Factgraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253.
- Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. Gpt4geo: How a language model sees the world’s geography. *arXiv preprint arXiv:2306.00020*.
- Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. 2024. Grapheval: A knowledge-graph based llm hallucination evaluation framework. *arXiv preprint arXiv:2407.10793*.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.
- Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. 2023. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2024. Fundamental limitations of alignment in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 53079–53112.
- Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024.  $\beta$ -dpo: Direct preference optimization with dynamic  $\beta$ . In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
- Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pages 4006–4017.
- Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.
- Yichi Zhang, Zhuo Chen, Yin Fang, Yanxi Lu, Li Fangming, Wen Zhang, and Huajun Chen. 2024. Knowledgeable preference alignment for LLMs in domain-specific question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 891–904, Bangkok, Thailand. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

## A Appendix

### A.1 Additional Clarification about Geospatial Hallucination’s Denotation and Connotation

We define geospatial hallucination of not only fabrication of non existing entities (relations), but also omission of actual things and confusion of their attributions. From this definition, geospatial hallucination is a subset of hallucination, which can be classified as a subset of ‘mistakes’ of LLMs.

LLMs’ mistake is a broader topic of models generating unsatisfactory contents. There are other kinds of "mistakes" like instruction following violation (Zhou et al., 2023), harmful content generation (Kumar et al., 2023), or misaligned values (Wolf et al., 2024), which are incorrect behaviors of LLMs even with no factual or logical error.

### A.2 Additional Explanation of DynamicKTO’s Motivation and Advantages

In this section, we would make extra theoretical explanations to our motivation of DynamicKTO from a model’s loss perspective.

To the best of our knowledge, we are the first to propose a method of adjusting  $\beta$  in KTO and demonstrate its advantages in geospatial hallucination mitigation.

We conducted analysis with a proposed metric called  $\mathcal{L}_{\text{FactDistance}}$  in the following definition, which is inspired by DPO and KTO loss as a way to assess model’s judgement between hallucinated and factual information.

Let  $\theta$  be the policy (model). Given a pair regarding a same knowledge, their log probabilities are defined as follows:

$$\log p_{\theta}(\mathbf{x}) = \frac{1}{L} \sum_{i=1}^L \log p_{\theta}(x_i \mid \text{prompt}, x_{<i}), \quad (1)$$

$$\log p_{\theta}(\mathbf{y}) = \frac{1}{L} \sum_{i=1}^L \log p_{\theta}(y_i \mid \text{prompt}, y_{<i}). \quad (2)$$

Here,  $\mathbf{x} = (x_1, \dots, x_L)$  denotes the factual sequence, and  $\mathbf{y} = (y_1, \dots, y_L)$  denotes the hallucinated sequence.

$$z^{(n)} = \log p_{\theta}(\mathbf{x}^{(n)}) - \log p_{\theta}(\mathbf{y}^{(n)}), \quad (3)$$



$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad (4)$$

$$\mathcal{L}_{\text{FactDistance}} = -\frac{1}{N} \sum_{n=1}^N \log \sigma(z^{(n)}). \quad (5)$$

The loss quantifies factual preference by comparing the log probabilities of a factual response against a hallucinated response.

We use the definition above to analyze the influences of training methods to model. The experiment results are shown in Table 7. Statistics of  $\mathcal{L}_{\text{FactDistance}}$  is demonstrated in Table 8.

Table 7:  $\mathcal{L}_{\text{FactDistance}}$  changes after training. Base model is the chat model instruction-tuned by its developer.

Category	Llama3.1-8B	+DPO	+KTO	+DynamicKTO
Entity	1.1072	1.0874	1.0770	0.7209
Relation	0.9465	0.9372	0.9278	0.8200
Attribute	0.7471	0.7401	0.7383	0.6265

Table 8: Training with DynamicKTO lowers both the mean and the variability (standard deviation) of  $\mathcal{L}_{\text{FactDistance}}$ .

Metric	Llama3.1-8B	+DPO	+KTO	+DynamicKTO
Macro Average	0.9336	0.9216	0.9144	0.7225
Standard Deviation	0.1472	0.1422	0.1386	0.0790

The difference of  $\mathcal{L}_{\text{FactDistance}}$  among Entity, Relation, and Attribute indicate a difference of difficulty for a LLM to judge and generate factual contents. As a result, a fixed  $\beta$  is not appropriate since it enforce same risk management strategy throughout all the data, which calls a dynamic adjusting  $\beta$  at task level (Entity, Relation, Attribute). We hope to 1) reduce the overall FactDistanceLoss and 2) enhance stability and consistency.

First, the introduction of the DynamicKTO optimization algorithm leads to a substantial improvement in factual preference, as evidenced by lower FactDistanceLoss scores across all three categories: Entity, Relation, and Attribute.

Second, the standard deviation measures the consistency of the model’s performance across different data points. The baseline model has a standard deviation of 0.1472, which decreases with DPO (0.1422) and KTO (0.1386). However, DynamicKTO achieves the most significant reduction to 0.0790, indicating that it not only improves factual preference but also enhances the stability

and consistency of the model’s responses across samples.

### A.3 Geospatial Hallucination Results on Underrepresented Regions

Geospatial hallucination situations of underrepresented regions apart from major cities like Beijing (GDP per capita \$ 30,177), London (GDP per capita \$ 79,069) , or New York (GDP per capita \$119,932) are also worth attention.

We understand the inherent geographically biases of LLMs introduced by pre-training or post-training stages. The results of geospatial hallucination in large cities is present in Table 2. Even in most important metropolises, the challenge of hallucination is great.

We expand our benchmark to include three more underrepresented regions globally, which are Cairo (in Africa, GDP per capita \$8,847), Kabul (in Asia, GDP per capita is \$1,188), and Sucre (in South America, no Wikipedia GDP data, Bolivia’s GDP per capita is \$4,014).<sup>5</sup> According to various factors, these regions have relative fewer information resources on the Internet, leading to less attention especially in LLMs.

Quantitative evaluation results are shown in Table 9. The lower performance results demonstrate that LLMs are facing great challenges of geospatial hallucination especially in underrepresented regions. Compared to results in paper, especially the better performance on New York, we can observe a general bias in the world knowledge that LLMs possess now.

### A.4 Training Cost Analysis of DynamicKTO

Through quantitative experiments and complexity analysis, the increase in training cost is manageable. While the introduction of additional operations in DynamicKTO theoretically slows down training, these operations are not computationally intensive. As a result, the impact on training efficiency is acceptable at most of the time. Empirical results show that with 4 NVIDIA A100 GPUs, standard KTO training takes 3h 1m 49s (181 mins), whereas DynamicKTO training takes 3h 7m 52s (187 mins) —a slight increase of 3.3%. In addition, since the additional operation is conducted once on each training sample, the training time scales linearly with the amount of data. The approach remains efficient for larger datasets. The slight increase in time cost is

<sup>5</sup>All GDP data above are from the Wikipedia entry of List of cities by GDP or the entry of Economy of Bolivia.

Table 9: Representative results on Cairo, Kabul, and Sucre. All the LLMs are chat models instruction-tuned by their developers. Results are sorted by model name.

Model	Cairo				Kabul				Sucre			
	Entity	Relation	Attribute	Overall	Entity	Relation	Attribute	Overall	Entity	Relation	Attribute	Overall
DeepSeek-V3	0.4883	0.3304	0.4440	0.4209	0.4817	0.3360	0.4280	0.4152	0.4900	0.3328	0.5000	0.4409
GPT-4o-mini	0.4917	0.3904	0.4040	0.4287	0.5600	0.3960	0.4400	0.4653	0.5533	0.4048	0.4760	0.4780
Llama-3.3-70B	0.4900	0.3840	0.4640	0.4460	0.5983	0.3632	0.4360	0.4658	0.5250	0.3672	0.4720	0.4547
Llama-3.1-70B	0.4717	0.3568	0.4080	0.4122	0.5583	0.3680	0.4240	0.4501	0.5200	0.3656	0.4400	0.4419
Llama-3.1-8B	0.4150	0.3696	0.2400	0.3415	0.3500	0.2984	0.2240	0.2908	0.4000	0.3704	0.2480	0.3395
Qwen2.5-72B	0.4367	0.3384	0.4640	0.4130	0.4067	0.3216	0.4600	0.3961	0.4617	0.3336	0.4960	0.4304
Qwen2.5-7B	0.3900	0.2778	0.4160	0.3613	0.3450	0.2989	0.4120	0.3520	0.4183	0.2834	0.4160	0.3726
Random	0.4367	0.4024	0.2680	0.3690	0.4167	0.3867	0.2040	0.3358	0.3933	0.3752	0.2720	0.3468

almost negligible considering its effectiveness in mitigating hallucinations.

### A.5 Additional Analysis about DynamicKTO’s Performance

DynamicKTO’s performance on geospatial hallucination mitigation is demonstrate in Table 5. In this section, we would conduct more analysis.

As a novel and competitive alignment method, ORPO is a strong baseline in specific settings. Putting aside the greater data efficiency of KTO, theoretical analysis suggests that if the preference data has sufficiently little noise and sufficiently little intransitivity, then KTO might fall behind of other alignment methods (Ethayarajh et al., 2024). ORPO is also a advanced algorithm featuring effectiveness, efficiency and scalability (Hong et al., 2024). The used factuality alignment dataset, especially the relation subset, is carefully curated, which may bring slight disadvantages for KTO as for relation category (about 1% accuracy difference). Generally, DynamicKTO is superior than ORPO on the task of geospatial hallucination mitigation. Considering its theoretical potential and empirical superiority, DynamicKTO is especially useful for hallucination mitigation.

### A.6 CityEval Performance Comparision between DynamicKTO and Baselines

DynamicKTO generalize better to CityEval benchmark than baselines. With the importation of additional finetuning data independent from CityEval tasks, it is expected for factuality aligned models to perform worse than CityGPT. Empirical results demonstrate that DynamicKTO can maintain the most comprehensive capabilities for CityEval while achieving SOTA geospatial hallucination mitigation as demonstrated in Table 10.

Table 10: DynamicKTO can generalize better to CityEval benchmark than other baselines. CI refers to City Image, US to Urban Semantics, SRR to Spatial Reasoning Route, and SRNR to Spatial Reasoning NoRoute. +DynamicKTO refers to Factual-CityGPT-Llama3.1-8B, which is finetuned on CityGPT-Llama3.1-8B with DynamicKTO.

Model	CityEval			
	CI	US	SRR	SRNR
CityGPT-Llama3.1-8B	0.5492	<b>0.7000</b>	<b>0.8440</b>	<b>0.6460</b>
+KTO	0.4077	0.5233	0.7880	0.6280
+DPO	0.4338	0.5567	0.7840	0.6200
+DynamicKTO	<b>0.5569</b>	0.6933	0.8040	0.6160

### A.7 Additional Analysis about DynamicKTO’s Difference from Sequential Training

Previous studies (Lee et al., 2024; Blakeney et al., 2024; Leybzoon and Kervadec, 2024) have examined the impact of training data order in language model fine-tuning, highlighting the critical role of data sequencing. In the context of achieving balanced geospatial factual capabilities, replacing DynamicKTO with sequential KTO tuning per task may be suboptimal due to the risk of overfitting to earlier tasks or overwriting previously learned knowledge.

To further investigate this, we conducted experiments using sequential KTO training, applying the same best-performing  $\beta$  for each individual task. The results, as shown in Table 11, show that DynamicKTO outperforms naive sequential KTO fine-tuning, reinforcing its advantage in managing knowledge injection across heterogeneous geospatial tasks.

### A.8 Automated Taxonomy Initialization of DynamicKTO for Different Tasks

The core insight behind DynamicKTO originates from the observed heterogeneity in data difficulty across different knowledge categories. However, a fixed  $\beta$  means the same risk management strategy

Table 11: DynamicKTO cannot be replaced by naively model training with sequential KTO with different hyperparameters. ST denotes Sequentially Trained. MT denotes Mixed Trained.

Method	Entity	Relation	Attribute	Overall
ST w/ KTO	0.4567	0.4168	0.3960	0.4232
MT w/ KTO	0.4333	0.3912	0.3000	0.3748
DynamicKTO	<b>0.5717</b>	<b>0.4256</b>	<b>0.4600</b>	<b>0.4858</b>

throughout the datasets in KTO, which is not appropriate anymore. A better algorithm, adjusting its "tightness" according to training knowledge, is needed for more effective hallucination mitigation.

So, the setting of  $\beta$  would be better decided from the statistics of the training data used and domain knowledge. We introduce a metric called  $\mathcal{L}_{\text{FactDistance}}$ , which quantifies a model's ability to distinguish between factual and hallucinated information. Experimental results in Table 7 reveal clear differences in judgment difficulty across categories: Entity is the most challenging, followed by Relation, and Attribute is the easiest. Based on this observation,  $\beta$  is set inversely proportional to  $\mathcal{L}_{\text{FactDistance}}$ .

Originally, DynamicKTO is developed specifically to mitigate geospatial hallucination, the primary focus of this study. To that end, we also introduce a Spatial World Hallucination Taxonomy, grounded in both domain understanding and broad geospatial data analysis. The consistency between the taxonomy, the  $\mathcal{L}_{\text{FactDistance}}$  analysis, and DynamicKTO's empirical results further supports the soundness of this task design.

In terms of tuning granularity, we aim to strike a balance between interpretability, training stability, and data efficiency. Our theoretical and empirical analysis shows that the Entity-Relation-Attribute categorization provides meaningful geospatial distinctions while maintaining training stability.

For domains beyond geospatial knowledge, researchers may want to rely on domain knowledge to set  $\beta$  in DynamicKTO. However, within the framework of  $\mathcal{L}_{\text{FactDistance}}$  analysis and DynamicKTO, the  $\beta$  can be automatically and easily calculated at different levels, removing the need for predefined task types. Two practical and promising alternatives can be utilized.

- Sample-Level:  $\beta$  is calculated individually for each training sample based on its
- Unsupervised Clustering-Level: Clustering

with (e.g., via K-Means++) is first applied to the data, and  $\beta$  is then computed based on each cluster's average data characteristics.

We also validate these methods for geospatial hallucination. As shown in Table 12, experimental results demonstrate significant improvements compared with the baseline method. On the other hand, these two methods still underperform the category-level DynamicKTO. This is mainly due to the instability. So we still recommend using DynamicKTO along with our taxonomy for the geospatial knowledge problem.

Table 12: The predefined taxonomy is helpful for geospatial hallucination mitigation. The number of clusters is set to three for fair comparison. S. Level denotes Sample-Level. UC. Level denotes Unsupervised Clustering-Level. C. Level denotes Category-Level.

Method	Entity	Relation	Attribute	Overall
+KTO	0.4333	0.3912	0.3000	0.3748
+DynamicKTO (S. Level)	0.4400	0.4336	0.4160	0.4299
+DynamicKTO (UC. Level)	0.4600	<b>0.4730</b>	<b>0.4080</b>	<b>0.4467</b>
+DynamicKTO (C. Level)	<b>0.5717</b>	0.4256	<b>0.4600</b>	<b>0.4858</b>

## A.9 Additional Analysis about GEOHALUBENCH's Uniqueness and Advantages over Wikipedia-Based Datasets

GEOHALUBENCH distinguishes itself among existing Wikipedia-based factual detection datasets. While some shared knowledge may be tested, GEOHALUBENCH has several key distinctions that set it apart.

First, its taxonomy is specially designed for the scenario of geospatial hallucination, whereas other fact-check benchmarks involve general factual accuracy.

Second, unlike wikipedia-based factual datasets like FEVER (Thorne et al., 2018) or WiCE (Kamoi et al., 2023), which contain claims and labels (SUPPORTED, UNSUPPORTED, etc.), GEOHALUBENCH can attribute a test sample's response to a specific type of hallucination. This feature enhances the interpretability of hallucination detection.

Third, GEOHALUBENCH is automatically constructed from a knowledge graph with data from OpenStreetMap and Foursquare's Open Source Places, which are high-quality, updating geoinformation services. As a result, GeoHaluBench is competitive in terms of data quality, scalability, coverage, and quantity.

Additionally, with the increasing demand for LLMs to master real-world geospatial knowledge, a specialized and high-quality benchmark like GeoHaluBench is essential.

Besides, previous related works have examined the possibility of using Wikipedia for geographic problems, but they found limitations of Wikipedia compared to specialized map services like OpenStreetMap. For instance, GeoLM (Li et al., 2023b) points out that training with Wikipedia only solves partial challenges in geospatial grounding, as it only provides the linguistic context of a geo-entity with sentences describing history, demographics, climate, etc. The information about the geospatial neighbors of a geoentity is still missing. Another research named GeoLLM (Manvi et al., 2023) also utilizes map data from OpenStreetMap rather than information from wikipedia.

## A.10 Demonstration of the Dataset

### A.10.1 Additional Discussion about the Question Form of GEOHALUBENCH

Multi-choice questions answering serve as the practice to detect hallucination behaviors in GEOHALUBENCH. If a LLM choose a non-existing option like Silver Spoon Cafe in Figure 3, we can reasonably infer that the LLM 'believes' that Silver Spoon Cafe is an actual POI, demonstrating its vulnerability to hallucination. Besides, multi-choice question can serve as a practical form of interacting with LLMs for its usability, interpretability, and controllability, especially when users regard them as knowledge bases apart from chatbots. We also get inspired by existing hallucination evaluation benchmarks like TruthfulQA (Lin et al., 2022), REALTIMEQA (Kasai et al., 2023), Med-HAL (Pal et al., 2023), FACTOR (Muhlgay et al., 2024), etc. which utilize multi-choice QA to detect and analyze various kind of hallucinations, providing valuable assessment for LLMs.

### A.10.2 GEOHALUBENCH Dataset

Figure 3 demonstrates a test case from Entity-POI-Existence category. The original options may contain Chinese characters, as the selected region is Beijing, China; these have been translated into English for demonstration purposes. The situation is similar in other non-English-speaking regions.

### A.10.3 DynamicKTO Fine-tuning Dataset

Figure 6 presents two examples used during fine-tuning with DynamicKTO. Although both examples

Figure 6: A positive and negative training sample used by DynamicKTO.

POSITIVE SAMPLE	
<b>TASK:</b> [POI_Category]	
<b>USER:</b> What category does the following POI (Point of Interest) in Beijing belong to: Ajisen Ramen Shuangjing Restaurant?	
<b>ASSISTANT:</b> Dining and Drinking > Restaurant > Asian Restaurant > Noodle Restaurant	
<b>Label:</b> Factual	
NEGATIVE SAMPLE	
<b>TASK:</b> [POI_Category]	
<b>USER:</b> What category does the following POI (Point of Interest) in Beijing belong to: Ajisen Ramen Shuangjing Restaurant?	
<b>ASSISTANT:</b> Dining and Drinking > Restaurant > Fast Food Restaurant	
<b>Label:</b> Hallucinated	

address the same question here, preference pairs are not required. A task label for hyperparameter adjustment, a content label for optimization, and a direction label are sufficient for DynamicKTO.

## A.11 Datasets and Baselines Detail for DynamicKTO Experiments

**Datasets.** We construct a fine-tuning dataset consisting of 1500 (for entity information) or 2000 (for relation or attribute information) instances of POI, AOI, and Road respectively. These elements are extracted randomly from *SpatialKG* and then organized into natural language narratives with templates. From the constructing process, they are annotated to be either hallucinated or factual naturally. In an attempt to generalize, a narrative is paraphrased. To avoid data leakage, none of the instances in the training set is identical to any sample in GEOHALUBENCH.

**Baselines.** We compare DynamicKTO with supervised fine-tuning, Direct Preference Optimization (DPO) (Rafailov et al., 2023), Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024), Simple Preference Optimization (SimPO) (Meng et al., 2025), and Odds Ratio Preference Optimization algorithm (ORPO) (Hong et al., 2024) for their effects to hallucination mitigation. For SFT, we refer the data and training pipeline from CityGPT (Feng et al., 2024a), a previous study of injecting LLM with urban knowledge. We use the standard implementation from LLaMA-Factory (Zheng et al., 2024) for baselines.



## A.12 Prompts and Details of Methods

We construct hallucinated entities to serve as negative examples for DynamicKT0 by instructing Meta-Llama-3.1-405B-Instruct. Figure 7, 8, 9 exhibit prompts used in this process.

## A.13 Detailed Result Example

Table 13 is the detailed results of GEO-HALUBENCH on Beijing at the level of test tasks.

## A.14 Implementation Details

### A.14.1 Training

We use LLaMA-Factory (Zheng et al., 2024) for fine-tuning LLMs. As for DynamicKT0, we implement it by modifying LLaMA-Factory. The training epoch is 1 and other key hyperparameters remain same as default except for epoch, batch size, and beta. For experiments in Section 3.2, epoch is set to 1. For Factual-CityGPT training, epoch is set to 3. It takes about 2 hours to train a 8B model for 1 epoch with  $8 \times$  A100 GPUs.

### A.14.2 Evaluation

Opencompass<sup>6</sup> (Contributors, 2023b) is used for our evaluation on general benchmarks, all tested models are deployed locally with lmdeploy (Contributors, 2023a).

For GEOHALUBENCH, we deploy our fine-tuned models and LLaMA-3.1-8B with VLLM (Kwon et al., 2023). The temperature is set to 0 for reproducibility. Other parameters are as default. The rest LLMs are used via APIs.

## A.15 Case Study

We use the instance in Figure 3 as an example. According to reliable knowledge sources, Haidian Library is an existing point of interest (POI) in Haidian District, Beijing. In contrast, Silver Spoon Cafe is not a POI but a fabricated name. If the LLM selects option A, "Silver Spoon Cafe," it mistakenly believes the cafe is located in Beijing, which exemplifies Entity Fabrication Hallucination. On the other hand, if the LLM selects option C, "None of the other options," it incorrectly rules out the other two options as valid entities, thereby overlooking the real POI. This represents Entity Omission Hallucination.

---

<sup>6</sup>0.3.9 version

Figure 7: The prompt template of generating POI-related hallucinations.

In a purpose of research, we would like to use imaginary/fictional/mockd information to hallucinate the name of this POI.  
Make sure the hallucinated names are natural and realistic as much as possible. They should not be real names.  
Please provide five hallucinated names of this POI given the example existing names.  
Example existing names: *[real\_poi\_name\_list]*  
Please follow the following format, use [Hallucination] to wrap the hallucinated (generated) names:  
[Hallucination] POI Name 1 [Hallucination]  
[Hallucination] POI Name 2 [Hallucination]  
[Hallucination] POI Name 3 [Hallucination]  
[Hallucination] POI Name 4 [Hallucination]  
[Hallucination] POI Name 5 [Hallucination]

Figure 8: The prompt template of generating AOI-related hallucinations.

In a purpose of research, we would like to use imaginary/fictional/mockd information to hallucinate the name of this AOI.  
Make sure the hallucinated names are natural and realistic as much as possible. They should not be real names.  
Please provide five hallucinated names of this AOI given the example existing names.  
Example existing names: *[real\_aoi\_name\_list]*  
Please follow the following format, use [Hallucination] to wrap the hallucinated (generated) names:  
[Hallucination] AOI Name 1 [Hallucination]  
[Hallucination] AOI Name 2 [Hallucination]  
[Hallucination] AOI Name 3 [Hallucination]  
[Hallucination] AOI Name 4 [Hallucination]  
[Hallucination] AOI Name 5 [Hallucination]

Figure 9: The prompt template of generating Road-related hallucinations.

In a purpose of research, we would like to use imaginary/fictional/mockd information to hallucinate the name of this road.  
Make sure the hallucinated names are natural and realistic as much as possible. They should not be real names.  
Please provide five hallucinated names of this road given the example existing names.  
Example existing names: *[real\_road\_name\_list]*  
Please follow the following format, use [Hallucination] to wrap the hallucinated (generated) names:  
[Hallucination] Road Name 1 [Hallucination]  
[Hallucination] Road Name 2 [Hallucination]  
[Hallucination] Road Name 3 [Hallucination]  
[Hallucination] Road Name 4 [Hallucination]  
[Hallucination] Road Name 5 [Hallucination]

Table 13: Detailed results of GEOHALUBENCH on Beijing. PE refers to POI-Existence, AE refers to AOI-Existence, RE refers to Road-Existence, PLoA refers to POI-LocateAt-AOI, PNeP refers to POI-Near-POI, ANeA refers to AOI-Near-AOI, ACoR refers to AOI-ConnectTo-Road, RCoR refers to Road-ConnectTo-Road, PAddr refers to POI-Address, PCate refers to POI-Category, ALand refers to AOI-LandUse, AArea refers to AOI-Area, RLeng refers to Road-Length. Models are in original names in APIs.

Model	PE	AE	RE	PLoA	PNeP	ANeA	ACoR	RCoR	PAddr	PCate	ALand	AArea	RLeng
DeepSeek-V3	0.180	0.255	0.600	0.332	0.084	0.024	0.260	0.368	0.200	0.800	0.520	0.040	0.100
gemini-2.0-flash	0.170	0.335	0.640	0.360	0.076	0.196	0.388	0.516	0.160	0.820	0.600	0.240	0.380
gpt-4o	0.135	0.290	0.505	0.320	0.104	0.116	0.268	0.308	0.160	0.680	0.460	0.080	0.200
gpt-4o-mini	0.270	0.390	0.485	0.360	0.124	0.144	0.324	0.264	0.140	0.760	0.540	0.040	0.200
Llama-3.1-8B-Instruct	0.385	0.375	0.455	0.452	0.320	0.240	0.408	0.392	0.040	0.640	0.500	0.200	0.040
Llama-3.1-8B-Instruct	0.385	0.375	0.455	0.452	0.320	0.240	0.408	0.392	0.040	0.640	0.500	0.200	0.040
Llama-3.3-70B-Instruct	0.320	0.345	0.475	0.476	0.196	0.108	0.300	0.276	0.180	0.840	0.580	0.020	0.120
Meta-Llama-3.1-405B-Instruct	0.160	0.265	0.510	0.324	0.096	0.140	0.252	0.160	0.100	0.840	0.600	0.000	0.000
Meta-Llama-3.1-70B-Instruct	0.290	0.330	0.460	0.468	0.232	0.276	0.328	0.204	0.280	0.800	0.640	0.000	0.040
Mistral-Small-24B-Instruct-2501	0.030	0.075	0.095	0.008	0.000	0.000	0.000	0.000	0.000	0.360	0.120	0.000	0.000
phi-4	0.225	0.325	0.475	0.352	0.068	0.128	0.368	0.320	0.200	0.800	0.620	0.080	0.200
qwen2.5-0.5b-instruct	0.395	0.465	0.475	0.484	0.444	0.372	0.496	0.480	0.160	0.600	0.480	0.320	0.260
qwen2.5-1.5b-instruct	0.235	0.240	0.515	0.440	0.184	0.156	0.348	0.352	0.320	0.760	0.480	0.180	0.300
Qwen2.5-14B-Instruct	0.055	0.105	0.365	0.132	0.004	0.000	0.088	0.032	0.120	0.700	0.220	0.000	0.000
Qwen2.5-32B-Instruct	0.085	0.165	0.465	0.164	0.020	0.008	0.136	0.156	0.200	0.760	0.420	0.000	0.040
qwen2.5-3b-instruct	0.165	0.300	0.450	0.292	0.032	0.008	0.040	0.064	0.020	0.740	0.180	0.000	0.000
Qwen2.5-72B-Instruct	0.085	0.180	0.480	0.184	0.024	0.028	0.140	0.192	0.120	0.780	0.420	0.020	0.020
Qwen2.5-7B-Instruct	0.125	0.160	0.405	0.236	0.020	0.008	0.092	0.064	0.100	0.700	0.260	0.100	0.160
qwen-max-2025-01-25	0.165	0.300	0.575	0.292	0.068	0.152	0.412	0.336	0.200	0.860	0.480	0.020	0.040
qwen-plus-2025-01-25	0.115	0.245	0.555	0.240	0.024	0.128	0.216	0.304	0.220	0.800	0.560	0.100	0.100
Random	0.290	0.285	0.315	0.220	0.324	0.244	0.296	0.304	0.240	0.260	0.240	0.200	0.220